Noisy Parallel Data Alignment

Ruoyu Xie, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{rxie, antonis}@gmu.edu

Abstract

An ongoing challenge in current natural language processing is how its major advancements tend to disproportionately favor resource-rich languages, leaving a significant number of under-resourced languages behind. Due to the lack of resources required to train and evaluate models, most modern language technologies are either nonexistent or unreliable to process endangered, local, and non-standardized languages. Optical character recognition (OCR) is often used to convert endangered language documents into machine-readable data. However, such OCR output is typically noisy, and most word alignment models are not built to work under such noisy conditions. In this work, we study the existing word-level alignment models under noisy settings and aim to make them more robust to noisy data. Our noise simulation and structural biasing method, tested on multiple language pairs, manages to reduce the alignment error rate on a state-of-the-art neural-based alignment model up to 59.6%.[1]

1 Introduction

Modern optical character recognition (OCR) software achieves good performance on documents in high-resource standardized languages, producing machine-readable text which can be used for many downstream natural language processing (NLP) tasks and various applications (Ignat et al., 2022; Van Strien et al., 2020; Amrhein and Clematide, 2018). However, attaining the same level of quality for texts in less-resourced local and non-standardized languages remains an open problem (Rijhwani et al., 2020).

The promise of OCR is particularly appealing for endangered languages, for which material might exist in non-machine-readable formats, such as physical books or educational materials. Digitizing such material could lead to the creation of NLP technologies for such otherwise severely under-served communities (Bustamante et al., 2020).

[1]Data and code are available online: https://github.com/ruoyuxie/noisy_parallel_data_alignment

to protinò ppetì tôχi ' aforammèna ? ”
[Clean text]

to protino ppeti toxi ’ aforamniena ? ’
[OCRed text]

to protino ppeti toxi ` aforammena ? ’
[Synthetic text]

Figure 1: Synthetic data example in Griko with character differences highlighted. Our synthetic data manage to mimic the real OCR noise.

Beyond the primary goal of digitizing printed material in endangered languages, the need for robust alignment tools is wider. The majority of the world's languages are being traditionally oral (Bird, 2020), which implies that to obtain textual data at scale one would need to rely on automatic speech recognition (ASR), which in turn would produce invariably noisy outputs. It is worth noting that the availability of translations can significantly improve systems beyond machine translation (MT), such as OCR (Rijhwani et al., 2020) or ASR (Anastasopoulos and Chiang, 2018). This creates a chicken-and-egg situation: on one hand, OCR and ASR can be used to obtain noisy parallel data; on the other hand, having good quality aligned data can improve OCR or ASR.

In this vein, We focus on the scenario of digitizing texts in a less-resourced language along with their translations (usually high-resource and/or widely spoken) similar to Rijhwani et al. (2020). Digitizing parallel documents can also be beneficial for educational purposes, as one could then create dictionaries through word- and phrase-level alignments, or ground language learning on another language (a learner's either L1 or L2). Also, as Ig-

Findings of the Association for Computational Linguistics: EACL 2023, pages 1501–1513
May 2-6, 2023 ©2023 Association for Computational Linguistics

nat et al. (2022) showed in recent work, such parallel corpora can be meaningfully used to create MT systems. However, the process that transforms digitized books or dictionaries into parallel sentences for training MT systems requires painstaking manual intervention.

In theory, the process could be semi-automated using sentence alignment methods, but in practice, the situation is very different: OCR systems tend to generate very noisy text for endangered languages (Alpert-Abrams, 2016, *inter alia*), which in turn leads to poor alignments between two parallel sides. As we show, alignment tools are particularly brittle in the presence of noise.

In this work, we take the first step towards solving the above issue. We investigate the relationship between OCR noise and alignment results and build a probabilistic model to simulate OCR errors and create realistic OCR-like synthetic data. We also manually annotate a total of 4,101 gold alignments for an endangered language pair, Griko-Italian, in order to evaluate our methods in a real-world setting. We leverage structural knowledge and augmented data, greatly reducing the alignment error rate (AER) for all four high- and low-resource language pairs up to 59.6%.

2 Problem Setting

Our work is a straightforward extension of previous word-level alignment work. Given a sequence of words $\mathbf{x} = (x_1, \ldots, x_n)$ in a source language and $\mathbf{y} = (y_1, \ldots, y_m)$ in a target language, the alignment model produces alignment pairs:[2]

$$\mathcal{A} = \{(x_i, y_j) : x_i \in \mathbf{x}, y_j \in \mathbf{y}\}$$

The difference with previous work is that the starting data will be the output of an OCR pipeline, hence producing noisy parallel data $(\mathbf{x}^*, \mathbf{y}^*)$ instead of "clean" $(\mathbf{x}, \mathbf{y})$ ones. The level of noise may vary between the two sides.

Hence, our goal is to produce an alignment

$$\mathcal{A}^* = \{(x_i, y_j) : x_i \in \mathbf{x}^*, y_j \in \mathbf{y}^*\}$$

that will be as close to the alignment $\mathcal{A}$ that we would have obtained without the presence of noise. We measure model performance using the alignment error rate (AER; Och and Ney, 2003) against the gold alignments.[3]

[2]Sometimes denoted with a latent variable, but we use an equivalent notation for simplicity.

[3]Lower AER means a better alignment. More details on the metric in Appendix A.

3 Method

We create synthetic data that mimic OCR-like noise, that can be used to train/finetune alignment models. Our ***simple yet effective*** method mainly consists of *(i)* building probabilistic models based on edit distance measures and capturing real OCR errors; *(ii)* creating synthetic (noisy) OCR-like data by applying our error-introducing model on clean parallel data; *(iii)* training or finetuning alignment models on synthetic data.

3.1 OCR Error Modeling

Error types For OCRed text, different types of texts, languages, and corpora will lead to different error distributions. At the character level, there are generally three types of OCR errors: insertions, deletions, and substitutions. In most cases, deletions and substitutions are more common, with spurious insertions being rarer.

Noise model By comparing the OCRed text with its post-corrected version, we use Levenshtein distance to compute the edit distances and the probability distributions of edits/errors over the corpus with a straightforward count-based approach.

We treat deletion error as part of substitution error. Given a sequence of characters $x_i, \ldots, x_j$ from a clean corpus $\mathbf{x}$ and a sequence of characters $y_i, \ldots, y_j$ from its OCRed noisy version $\mathbf{y}$, we simply count the number of times a correct character x_i is recognized as character y_i (or recognized as the empty character ϵ if it is erroneously deleted). We can then compute the probability of an erroneous substitution or deletion as follows:

$$P_{sub}(x_i \to y_i) = \frac{\text{count}(x_i \to y_i)}{\text{count}(x_i)}$$

and the overall substitution error distribution is conditioned on the correct character x_i:

$$\mathcal{D}_{sub}(x_i) \sim P_{sub}(x_i \to y_i).$$

For insertion errors, we consider that insertion occurs when ϵ becomes another character y_i and count the number of times that insertion occurs after its previous character. A special token <begin> is used when insertion occurs at the beginning of the sentence. In general, we calculate the insertion error probability with:

$$P_{ins}(x_{i-1}\epsilon \to x_{i-1}y_i) = \frac{\text{count}(x_{i-1}\epsilon \to x_{i-1}y_i)}{\text{count}(x_{i-1}\epsilon)}$$

and the insertion error distribution for x_{i-1}:

$$\mathcal{D}_{ins}(x_{i-1}\epsilon) \sim P_{ins}(x_{i-1}\epsilon \rightarrow x_{i-1}y_i).$$

3.2 Data Augmentation

Synthetically noised data can be created by leveraging the calculated probability distributions from the previous section and traversing through the clean corpus for every character.

For each character **c**, we obtain its probability to be erroneous in the OCR output by sampling from the distribution of the substitution and insertion probabilities $\mathcal{D}_{ins}(c), \mathcal{D}_{sub}(c)$.[4] We randomly decide whether to add an error here based on its error distribution.

If an error will be introduced on **c**, we then randomly choose its corresponding error based on $P_{sub}(c)$ or $P_{ins}(c)$ depending on either substitution or insertion operation receptively.

Our method attempts to mimic the real OCR errors in given languages and corpus, resulting in very similar noise distributions. Figure 1 shows a side-by-side comparison of three versions of the same sentence, to showcase how realistic our synthetic text is.

3.3 Model Improvement

Given our synthetically noised parallel data, and potentially along with the original clean parallel data, we can now train or finetune a word alignment model to improve the model performance.

In the case of unsupervised models like the IBM translation models (Brown et al., 1993), `fast-align` (Dyer et al., 2013), or `Giza++` (Och, 2003), we simply train on the concatenation of all available data.

We also work with the state-of-the-art neural alignment model of Dou and Neubig (2021), which is based on Multilingual BERT (`mBERT`) (Devlin et al., 2019).[5] For this model, we distinguish two cases: supervised and unsupervised finetuning.[6] Under a supervised setting, we first obtain silver alignments from the clean dataset and use them as targets for the synthetic noisy data. The unsupervised setting is conceptually similar to training models like `Giza++`: we feed synthetically-noised sentence pairs into the alignment model, without using the target alignment as supervision. In low-resource scenarios, we leverage a diagonal bias to further improve the model's performance.

Language	Total CER	Sub. %	Ins. %
English	7.4	79	21
German	4.9	87	13
French	4.8	85.7	14.3
Griko	3.3	96.8	3.2
Ainu	1.4	91.9	8.1

Table 1: Total character error rate (CER) and percentage of substitution and insertion errors. Generally, substitution is the most common error in OCR output.

4 Languages and Datasets

We study four language pairs with varying amounts of data availability: English-French, English-German, Griko-Italian, and Ainu-Japanese.[7]

4.1 Dataset for Error Extraction

The ICDAR 2019 Competition on Post-OCR Text Correction (Rigaud et al., 2019) dataset provides both clean and OCRed text for English-French and English-German, which we use our noisy model to learn and mimic OCR errors for English, French, and German.

For Griko-Italian and Ainu-Japanese, Rijhwani et al. (2020) provide around 800 OCRed noisy and clean (post-corrected) sentences for both Griko and Ainu, from which we extract error distributions; for Italian and Japanese, only OCRed text is provided.[8]

To understand the characteristics of our datasets, we report the observed CER in Table 1. Generally, substitutions are the most common errors. Notice that Griko and Ainu have seemingly lower scores than any high-resource languages; that's because both use the Latin alphabet, the data that were digitized are typed in books with high-quality scans.[9]

4.2 Synthetic Data

We create synthetic data by applying captured OCR noise on clean text. For English, French, and German, the clean text comes from Europarl v8 corpus (Koehn, 2005). For Ainu, there are 816 clean sentences from Rijhwani et al. (2020), from which

[4] Including a third option for not inserting an error

[5] See Section 5.1 for more details.

[6] We use provided default parameters for both cases, which can be found on https://github.com/neulab/awesome-align

[7] Griko and Ainu are both under-resourced endangered languages.

[8] The quality of the OCR model on these high-resource languages are generally reliable.

[9] The English, French, and German data from ICDAR have lower-quality scans.

Language	Real CER	Syn. CER	Diff.
English	7.4	6.5	0.9
German	4.9	6.7	1.8
French	4.8	5.3	0.5
Griko	3.3	3.3	0
Ainu	1.4	1.0	0.4

Table 2: Our synthetically-noisy data have similar CER compared to the real OCR outputs, which implies that the real OCRed noisy data can be mimicked by our noise simulation model.

Model	Clean	OCRed	Diff.
`IBM 1`	43.7	49.2	5.5
`IBM 2`	37.3	43.4	6.1
`Giza++`	**14.5**	**20.8**	6.3
`fast-align`	19.8	25.7	5.9
`awesome-align`	45.1	48.8	**3.7**

Table 3: AER comparison for Griko-Italian. `Giza++` performs best on both settings, but it exhibits the largest drop in performance.

we keep the first 300 lines as test set and use the rest to create synthetic data. Anastasopoulos et al. (2018) provide 10,009 clean sentences for Griko. Table 2 shows the CER comparison between our synthetic data and real OCR data.

4.3 Test Set and Gold Alignment

The test set and gold alignment for English-French come from Mihalcea and Pedersen (2003). For English-German, the test set and gold alignments come from Europarl v7 corpus (Koehn, 2005) and Vilar et al. (2006), respectively. To study the effect of OCR-like errors on alignment, we create synthetically-noised test sets for both languages pairs by applying noise on one side or both, which results in four copies of the same test set: clean-clean, clean-noisy, noisy-clean, and noisy-noisy.

For low-resource language pairs, Rijhwani et al. (2020) provide about 800 parallel sentence pairs for each. We use the first 300 sentence pairs as our test sets. For the purpose of fair evaluation in our method, we annotate a total of 4,101 *gold* word-level alignment pairs for Griko-Italian test set. On the other hand, we obtain *silver* alignments from `awesome-align` for Ainu-Japanese as there is no existing gold alignment data available.[10]

5 Experiments

In this section, we present multiple experiments and demonstrate that our method results in significant AER reductions.

5.1 Experimental Setup

Models We study the following models:

- `IBM model 1&2` (Brown et al., 1993): the classic statistical word alignment models. They underpinned many other statistical machine translation and word alignment models.
- `Giza++` (Och, 2003): a popular statistical alignment model that is based on a pipeline of IBM and Hidden Markov models (Vogel et al., 1996).
- `fast-align` (Dyer et al., 2013): a simple but effective statistical word alignment model that is based on IBM Model 2, with an additional bias towards monotone alignment.
- `awesome-align` (Dou and Neubig, 2021): a neural word alignment model based on `mBERT`. It fine-tunes a pre-trained multilingual language model with parallel text and extracts the alignments from the resulting representations.

5.2 The Effect of OCR-like Noise

We use Griko-Italian as our main evaluation pair due to the presence of its gold alignments, which can most accurately reflect the model's performance under a low-resource scenario.

We first benchmark model performance on clean and OCRed parallel text to quantify OCR-error effects on alignment (Table 3). We compute AER for the clean and OCRed versions of Griko-Italian by comparing their alignment against our manually created gold alignment. We benchmark five different models that lead to several observations. First, note that clean text always results in a better alignment for all models. Overall, `Giza++` performs best among the models, but note that it also suffers the largest drop in performance when faced with noisy text. On the other hand, a vanilla `awesome-align`, which is otherwise a state-of-the-art model for languages that were included in the pre-training of its underlying model, performs the worst, not being better than a simple `IBM 1`.

We can thus conclude that OCR error does impact alignment quality for both statistical and neural based alignment models.

It is of note that for Griko-Italian every statistical model outperforms `awesome-align` in almost all cases. We hypothesize that this is due to the lack

[10] While not ideal, we can still measure how different results are when comparing alignments on clean versus noisy data.

	Griko-Italian		Ainu-Japanese	
	Clean	OCRed	Clean	OCRed
BASE	45.1	48.8	28.2	29.2
UNSUP-FT (A)	23.2±1.6	28±1.1	21.1 ±1	22.2±1.2
SUP-FT (B)	22.3±2	26.6±1.1	30.9±2.1	31.5±1.2
+ structural bias				
UNSUP-FT (A)	18.7±1	24.2±0.6	**15.3±3.9**	**13.8±4.2**
SUP-FT (B)	**18.2±2.6**	**22.9±2.4**	26.3 ±2.1	27.4±1.8
AER reduction	59.6%	53.1%	45.7%	52.7%

Table 4: For both endangered languages, our approach greatly reduces AER for both clean and OCRed data.

of structural knowledge; we deal with this in Section 5.3.1. awesome-align's low performance can also be explained by the fact that Griko is not well supported by its underlying representation model: Griko was not part of the pre-training language mix, and it does not use the same script as its closest language that was included in pre-training (Greek),[11] an important factor according to Muller et al. (2021). Compared to statistical models, we also observe considerably fewer alignment pairs are produced by awesome-align (Appendix 8), which might also be a contributing factor.

5.3 Making awesome-align Robust

The performance of awesome-align raises an intriguing question - Is the state-of-the-art neural based model capable to align noisy text, especially from low-resource languages. Given its general higher performance on many popular languages (Dou and Neubig, 2021) and the stability between clean and noisy text,[12] we make awesome-align as our main experiment target.

5.3.1 Low-Resource Setting

We introduce structural bias and propose two models: model (A) and model (B) finetuned in unsupervised and supervised settings respectively.

Structural Bias Structural alignment biases are widely used in statistical alignment models such as Brown et al. (1993); Vogel et al. (1996); Och (2003); Dyer et al. (2013). However, it is a missing component in awesome-align. Following by Dyer et al. (2013), we introduce diagonal bias and apply it on the top of awesome-align's attention layer. We create *(i)* a bias matrix M_b based on the position of the alignment, where the positions near the diagonal of the alignment matrix have the higher weights (See Figure 2); *(ii)* a tune-able hyper-parameter λ represents the weight of the bias. We set λ=1 for all low-resource language experiments; *(iii)* an average matrix M_{avg} that is the average of the original attention score, which is used for smoothing λ to make it where 1 represents maximum bias and 0 means no bias at all. We update the original awesome-align attention score A_{sc}:

$$A_{sc} = \lambda * A_{sc} + (1 - \lambda) * A_{sc} * M_b * M_{avg}$$

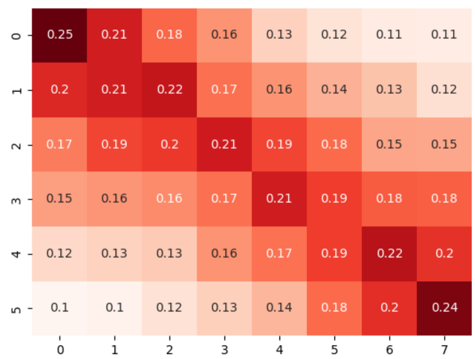

Figure 2: A sample 6 × 8 diagonal bias matrix. Darker color means stronger bias emphasis. We follow the same steps from Dyer et al. (2013) to calculate each position based on given rows and columns.

Our proposed models For our unsupervised-finetuned model (A), we create the synthetically-noised data by introducing OCR-like noise on clean parallel data, and then simply finetune the baseline model with all available data from both clean and synthetic text.

For the supervised-finetuned model (B), we first finetune an out-of-the-box awesome-align with the clean data from Anastasopoulos et al. (2018) and Rijhwani et al. (2020) for Griko-Italian and

[11] Modern Greek uses the Greek alphabet, while Griko uses the Latin alphabet.

[12] Lowest AER difference between clean and noisy text amount to all models.

Test Set	English-French Baseline	English-French Unsup-FT	English-French Sup-FT	English-French Reduction	English-German Baseline	English-German Unsup-FT	English-German Reduction
CLEAN-CLEAN	5.6	**4.6**	15.9	17.0%	17.9	**15.2**	15.1%
CLEAN-SYNTH	40.5	36.3	**29.4**	27.4%	43.8	**39.4**	10.0%
NOISY-CLEAN	39.2	34.2	**28.6**	27.0%	52.8	**50**	5.3%
NOISY-NOISY	53.6	46.1	**37.3**	30.4%	66.6	**63.5**	4.7%

Table 5: Result of `awesome-align` on English to French and German alignment. Both unsupervised and supervised finetuning with noise-induced data leads to big AER reduction when aligning noisy data. Reductions in German are less pronounced. Unsupervised finetuning with noisy data also improves clean-data alignment.

	`IBM 1` Clean	`IBM 1` OCRed	`IBM 2` Clean	`IBM 2` OCRed	`Giza++` Clean	`Giza++` OCRed	`fast-align` Clean	`fast-align` OCRed
Baseline	43.7	49.2	37.3	43.4	14.5	20.8	19.8	25.7
Train w. clean	**40.2**	**45.7**	**32.7**	**38.1**	**13.1**	**19.0**	**17.9**	**24.2**
Train w. noise	84.2	84.6	80.0	80.8	19.7	25.8	22.8	27.9

Table 6: Experiment on Griko-Italian, every statistical model benefits from training with additional clean data but suffers significant performance drops with synthetic noisy data, suggesting that traditional statistical models rely on clean text.

Aiun-Japanese respectively, which produces silver alignment. Next, we use the silver alignment as supervision to finetune `awesome-align` with synthetic noisy data.

We report the average plus-minus standard deviation of three runs for each model. Table 4 summarizes the results for our proposed models. We end up with around ***50%*** AER reduction for both endangered language pairs.

5.3.2 High-Resource Setting

We evaluate our data augmentation method on high-resource language pairs. Up to 400K synthetically noised English-French data was used for unsupervised finetuning. We also offer an additional reference data point, using 100K synthetic noised English-German data for unsupervised fine-tuning.

For supervised finetuning, we use up to 1M synthetic data. As before, we use silver alignments from clean data as supervision to finetune its synthetic noisy version, which does not require any additional human annotation effort.

Under both settings, model performance will plateau when adding more data. The results are summarized in Table 5. Both unsupervised and supervised finetunings with synthetically-noised data significantly improve alignment quality, especially for noisy test sets, in line with our previously presented results in low-resource settings.

5.4 Addtional Data on Statistical Models

We conduct additional experiments to find out whether training with additional data aids statistical models for endangered languages. We evaluate model performance on Griko-Italian.

We concatenate additional data to the examples comprising the test set. We first train the models with all 800 clean sentence pairs taken from Rijhwani et al. (2020) (which include the 300 sentences of the test set). Next, instead of using clean data, we substitute it with synthetically noised data and train the models.

The result is presented in Table 6. For every statistical model, training with additional clean text reduces AER. However, training with additional noisy text considerably hurts the models. The result shows that these statistical models rely on ***clean text*** to improve, which is almost always ***unavailable*** for endangered languages. This also implies that investing time in manually cleaning OCR data could be effective for these models; however, it is not always possible and contradicts the goal of reducing the human effort in this work.

6 Analysis and Discussion

In this section we conduct several analyses to better understand our method.

Incorporating Diagonal Bias As shown in Table 4, our diagonal bias markedly improves ev-

Test set	En-Fr	En-De
CLEAN-CLEAN	5.6	17.9
CLEAN-NOISY	40.5	43.8
NOISY-CLEAN	39.2	52.8
NOISY-NOISY	53.6	66.6

Table 7: `awesome-align` baseline on En-Fr and En-De. OCR-like noise dramatically degrades the performance.

ery test case for both endangered language pairs. Note that the attention score will be increased significantly by adding bias, which will still be a valid input for the final alignment matrix due to its alignment extraction mechanism (Dou and Neubig, 2021). In this work, we only apply diagonal bias under low-resource settings since it was shown in Dou and Neubig (2021) that growing heuristics such as grow-diag-final (Koehn et al., 2005; Och and Ney, 2000) do not achieve promising results for multiple high-resource language test sets.

Degradation of Alignment Table 7 presents the evaluation of four test sets for `awesome-align` in English-French and English-German. We observe a significant decline in performance when OCR-like noise is introduced. For example, with clean parallel text, the AER for English-French is 5.6%, but when OCR-like noise is added, the AER jumps to 53.6%, almost a tenfold increase.

Size of synthetic data We conduct quantitative analyses as shown in Figure 3 to examine `awesome-align` with different sizes of English-French synthetic data under both unsupervised and supervised settings. For space economy reasons, here we only discuss the results of the more challenging noisy-noisy test set. Note that dramatic degradation of alignment is observed when applying OCR-like noise to clean text (see Table 7). In general, the model produces better alignment as more data are used. However, there is also a tradeoff on the clean-clean test set as its performance worsens in the supervised scenario. Keep in mind, though, that this situation is only observed in high-resource language pairs; for a low-resource language pair like our Griko-Italian, in limited ablation experiments we found that we have not reached the data saturation point yet as more data simply resulted in better performance for both clean and noisy text.

Varying degrees of CER In a real-world scenario, the CER of OCRed data is typically unknown due to the absence of clean text. We investigate how different degrees of CER affect alignments by creating several English-French synthetic data with varying degrees of CER, testing them on `awesome-align`. We elaborate on the process and results in Appendix B.1. The main finding is that higher CER leads to greater AER, which is expected. However, we also find that mixing with different degrees of CER generally produces better results than a fixed CER throughout the corpus, suggesting that our augmentation approach could also work on the unknown CER real-world scenario.

Statistical Model vs Neural Model The question of which model to use in practical scenarios, though, remains tricky to answer. Due to similarities between Griko and Italian and prolonged language contact over centuries, the two languages follow very similar syntax; as a result, their alignment is largely monotone, which benefits models like `Giza++` and `fast-align`. They outperform, in fact, the vanilla neural `awesome-align` model by a large margin (see Table 3). However, this will not always be the case. For example, most books with parallel data in the Archive of Indigenous Languages of Latin America (AILLA) mostly contain data between indigenous languages and one of Spanish or English. Now the two sides of the data come from different language families and a monotone alignment is not necessarily to be expected. In such cases, it could indeed be the case that a more adaptable neural model like `awesome-align`, aided by our data augmentation and diagonal biasing methods, could indeed be the best option.

Different side of OCR noise An important insight derived from Table 5 is that the performance of `awesome-align` deteriorates significantly more when both sides of the parallel data are noisy, as compared to when only one side is noisy. This is in fact encouraging for our envisioned application scenarios, since, as in the AILLA examples described above, we expect that OCRed parallel data in endangered languages will come with one side in a high-resource standardized language like English and Spanish which in turn we expect the OCR model to be able to adequately handle.[13]

[13] Rijhwani et al. (2020) and Rijhwani et al. (2021) make similar observations on all endangered language datasets they work with.

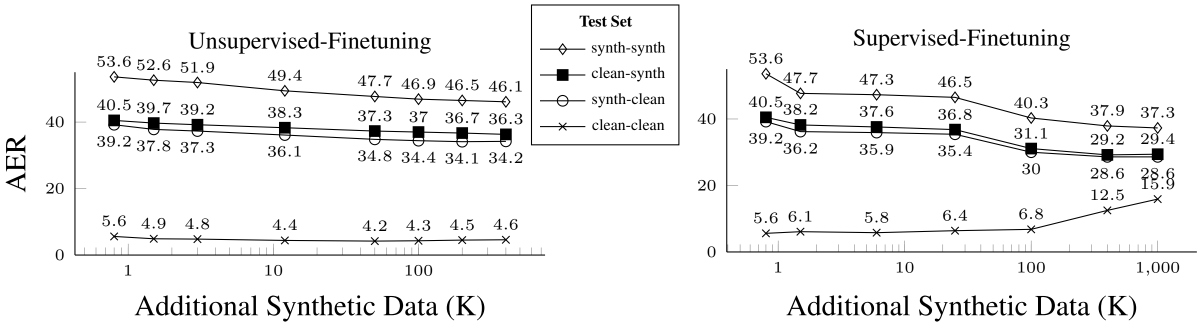

Figure 3: Ablation on English-French with varying degrees of additional synthetically noised data. Notice the log scale on the x-axis. The left-most point corresponds to no additional synthetic data (baseline). More data reduce AER for noisy test sets, especially in the supervised finetuning setting.

7 Related Work

Our work is a natural extension of previous word alignment work. A robust alignment tool for low-resource languages benefits MT systems (Xiang et al., 2010a; Levinboim and Chiang, 2015; Beloucif et al., 2016a; Nagata et al., 2020), or speech recognition (Anastasopoulos and Chiang, 2018), especially if sentence-level alignment tools like LASER (Artetxe and Schwenk, 2019; Chaudhary et al., 2019) do not cover all languages, so one may need to fall-back to word-level alignment heuristics to inform sentence-alignment models like `Hunalign` (Varga et al., 2007).

Research on word-level alignment started with statistical models, with the IBM Translation Models (Brown et al., 1993) serving as the foundation for many popular statistical word aligners (Och and Ney, 2000, 2003; Och, 2003; Tiedemann et al., 2016; Vogel et al., 1996; Och, 2003; Gao and Vogel, 2008; Dyer et al., 2013). In recent years, different neural network based alignment models gained in popularity including end-to-end based (Zenkel et al., 2020; Wu et al., 2022; Chen et al., 2021), MT-based (Chen et al., 2020), and pre-training based (Garg et al., 2019; Dou and Neubig, 2021). As `awesome-align` achieves the overall highest performance, we choose to focus on `awesome-align` in this work.

Some works involve improving word-level alignment for low-resource languages such as utilizing semantic information (Beloucif et al., 2016b; Pourdamghani et al., 2018), multi-task learning (Levinboim and Chiang, 2015), and combining complementary word alignments (Xiang et al., 2010b). None of the previous work, though, to our knowledge, tackles the problem of aligning data with OCR-like noise on one or both sides. The idea of augmenting training data is not new and has been applied in many areas and applications. Marton et al. (2009) augment data with paraphrases taken from other languages to improve low-resource language alignments. While potentially orthogonal to our approach, this idea is largely inapplicable to our endangered language settings, as we often have to work with the only available datasets for these particular languages. Applying structure alignment bias on statistical and neural models is also a well-studied area (Cohn et al., 2016; Brown et al., 1993; Vogel et al., 1996; Och, 2003; Dyer et al., 2013). However, to the best of our knowledge, we are the first to apply it to low-resource languages, proving that such an approach can greatly aid the real endangered language data.

8 Conclusion

In this work, we benchmark several popular word alignment models under OCR noisy settings with high- and low-resource language pairs, conducting several studies to investigate the relationship between OCR noise and alignment quality. We propose a simple yet effective approach to create realistic OCR-like synthetic data and make the state-of-the-art neural `awesome-align` model more robust by leveraging structural bias. Our work paves the way for future word-level alignment-related research on underrepresented languages. As part of this paper, we also release a total of 4,101 ground truth word alignment data for Griko-Italian, which can be a useful resource to investigate word- and sentence-level alignment techniques on practical endangered language scenarios.

9 Limitations

Using AER as the main evaluation metric could be a limitation of our work as it might be misleading in some cases (Fraser and Marcu, 2007). Another limitation, of course, is that we only manage to explore the tip of the iceberg given the sheer number of endangered languages. While we are confident in the results of both low-resource language pairs, our experiments on Ainu-Japanese could potentially lead to inaccurate AER since we use the automatically generated silver alignment. In the future, we hope to eventually annotate it with either the help of native speakers or dictionaries. We also plan to explore other alternative metrics and expand our alignment benchmark on as many endangered languages as possible.

Acknowledgements

We are thankful to Shruti Rijhwani and Graham Neubig, as well as the anonymous reviewers, for their valuable comments on the early stages of this work. We would also like to thank Sina Ahmadi for his useful feedback and the GMU Office of Research Computing for the computing resources. This work was supported by NEH Award PR-276810-21 as well as through a GMU OSCAR award for undergraduate research.

References

Hannah Alpert-Abrams. 2016. Machine reading the primeros libros. *Digital Humanities Quarterly*, 10(4).

Chantal Amrhein and Simon Clematide. 2018. Supervised ocr error detection and correction using statistical and neural machine translation methods. *Journal for Language Technology and Computational Linguistics (JLCL)*, 33(1):49–76.

Antonios Anastasopoulos and David Chiang. 2018. Leveraging translations for speech transcription in low-resource settings. *Proc. Interspeech 2018*, pages 1279–1283.

Antonios Anastasopoulos, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-speech tagging on an endangered language: a parallel griko-italian resource. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2529–2539.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Meriem Beloucif, Markus Saers, and Dekai Wu. 2016a. Improving word alignment for low resource languages using English monolingual SRL. In *Proceedings of the Sixth Workshop on Hybrid Approaches to Translation (HyTra6)*, pages 51–60, Osaka, Japan. The COLING 2016 Organizing Committee.

Meriem Beloucif, Markus Saers, and Dekai Wu. 2016b. Improving word alignment for low resource languages using english monolingual srl. In *Proceedings of the Sixth Workshop on Hybrid Approaches to Translation (HyTra6)*, pages 51–60.

Steven Bird. 2020. Sparse transcription. *Computational Linguistics*, 46(4):713–744.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from pdf files of truly low-resource languages in peru. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. *WMT 2019*, page 263.

Chi Chen, Maosong Sun, and Yang Liu. 2021. Mask-align: Self-supervised neural word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4781–4791, Online. Association for Computational Linguistics.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for Computational Linguistics.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022. OCR Improves Machine Translation for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Tomer Levinboim and David Chiang. 2015. Multi-task word alignment triangulation for low-resource languages. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1221–1226, Denver, Colorado. Association for Computational Linguistics.

Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Singapore. Association for Computational Linguistics.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, pages 1–10.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When Being Unseen from mBERT is just the Beginning: Handling New Languages with Multilingual Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462.

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, pages 160–167.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Nima Pourdamghani, Marjan Ghazvininejad, and Kevin Knight. 2018. Using word vectors to improve word alignments for low resource machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 524–528, New Orleans, Louisiana. Association for Computational Linguistics.

Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2019. ICDAR 2019 competition on post-OCR text correction. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1588–1593. IEEE.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 5931–5942, Online. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically aware semi-supervised learning for ocr post-correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

Jörg Tiedemann, Fabienne Cap, Jenna Kanerva, Filip Ginter, Sara Stymne, Robert Östling, and Marion Weller-Di Marco. 2016. Phrase-based SMT for Finnish with more data, better models and alternative alignment and translation tools. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 391–398, Berlin, Germany. Association for Computational Linguistics.

Daniel Van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of OCR quality on downstream NLP tasks. *SCITEPRESS-Science and Technology Publications*.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

David Vilar, Maja Popović, and Hermann Ney. 2006. AER: Do we need to "improve" our alignments? In *Proceedings of the Third International Workshop on Spoken Language Translation: Papers*.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Di Wu, Liang Ding, Shuo Yang, and Mingyang Li. 2022. Mirroralign: A super lightweight unsupervised word alignment model via cross-lingual contrastive learning. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 83–91.

Bing Xiang, Yonggang Deng, and Bowen Zhou. 2010a. Diversify and combine: Improving word alignment for machine translation on low-resource languages. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 22–26.

Bing Xiang, Yonggang Deng, and Bowen Zhou. 2010b. Diversify and combine: Improving word alignment for machine translation on low-resource languages. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 22–26, Uppsala, Sweden. Association for Computational Linguistics.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617, Online. Association for Computational Linguistics.

A Evaluation metric

We calculate precision, recall and alignment error rate as described in Och and Ney (2003), where A is a set of alignments to compare, S is a set of gold alignments, and P is the union of A and possible alignments in S. We then compute AER with:

$$\text{Precision} = \frac{|A \cap P|}{|A|} \quad \text{Recall} = \frac{|A \cap S|}{|S|}$$

$$\text{AER}(S, P; A) = 1 - \frac{|A \cap S + A \cap P|}{|A + S|}$$

B Additional Analyses

B.1 Varying degrees of CER

We create eight 100k English-French synthetic datasets with different CER for each: "unified" datasets with exactly the same CER on both sides: 2, 5, 10, and one with mixed CER with equally shared portions with 2, 5 and 10 CER; and "varying" datasets with slightly different CER between the English and French side, but in the same range as the others. We then finetune the model with these synthetic training datasets and compare against the "skyline" result presented before, where the augmentation matched the level of the true CER.

The results are presented in Table 9 for both the unsupervised- and the supervised-finetuning setting. Encouragingly, despite different CER in the augmentation data, there are no significant performance differences in most cases, especially for the unsupervised setting. Of course, levels of noise that match the true level tend to perform better or close to best overall. On the other hand, high levels of noise that lead to very high word error rate (WER)[14] cause a large degradation in the performance of the supervised finetuning approach, but do not seem to significantly affect the unsupervised approach.

Even more encouragingly, an augmented dataset that uses a mixture of different target CER (such as having a third of the dataset having a CER around 2, a third with CER around 5, and a third around 10 – named "mixed" in Table 9) in the supervised setting further outperforms the *informed skyline* which uses additional knowledge that might not be available (the true CER of the data to be aligned). For instance, in the clean-noisy test set this model reduces AER by a further 5% (from 31.1 to 29.3) and on the clean-clean test set it reduces AER by 19% (from 6.8 to 5.5). This means that our augmentation approach with varying levels of noise could be applied to any scenario, even if one does not know the level of noise present in the data-to-be-aligned.

[14]For example, a CER of around 10 translates to a WER of more than 70, meaning that (approximately) only 3 out of 10 words are correct.

	`IBM 1`		`IBM 2`		`Giza++`		`fast-align`		`awesome`	
	Clean	OCR	Clean	OCR	Clean	OCR	Clean	OCR	Clean	OCR
# of pairs	3844	3839	3833	3855	3810	3813	3801	3794	2978	2969
Precision	58.2	52.5	64.9	58.6	88.7	82.2	83.4	77.3	65.3	64
Recall	54.5	49.2	60.7	54.8	82.4	76.4	77.3	71.5	47.4	44.2

Table 8: Comparing the number of alignment pairs produced by models on Griko-Italian. `awesome-align` produces almost 25% less alignment pairs, resulting in markedly lower precision/recall and higher AER.

CER (WER) on Synthetic Data	Clean-Clean		Clean-Noisy		Noisy-Clean		Noisy-Noisy	
	UNSUP-FT	SUP-FT	UNSUP-FT	SUP-FT	UNSUP-FT	SUP-FT	UNSUP-FT	SUP-FT
Skyline: Using exactly the CER of the test set								
7.4-4.8 (59.7-51.7)	4.3	6.8	37	31.1	34.4	30	46.9	40.3
Unified: Exactly the same CER on both sides								
2-2 (32.2-29.1)	4.1	**5.1**	37.5	33.3	35.1	30.1	48.4	41.8
5-5 (55.1-52.4)	4.3	7.4	37.2	30.4	34.6	29.8	47.4	**40.0**
10-10 (72.2-71)	4.5	32.8	**36.8**	36.2	34.5	47.7	46.8	47.3
mixed (55-54.1)	5.4	5.5	37.3	39.6	35.3	39.8	47.7	54.2
Varying CER between the two parallel sides								
1.6-2.1 (27.8-29.6)	**4.0**	6.2	37.6	31.6	35	30.6	48.4	41.9
4.1-5.1 (49.6-52.5)	4.3	7.7	37	29.8	34.7	30.1	47.2	40.2
8.1-9.6 (66.9-69.5)	4.5	31.3	37.1	36.2	**34.3**	46.3	**46.7**	46.9
mixed (47.9-50)	4.2	5.6	37	**29.3**	34.6	**29.7**	47.2	40.4

Table 9: AER comparison for varying CER in 100K English-French augmented data used for either unsupervised or supervised finetuning. We highlight the best result under each setting and test set. Overall, most models' performance is close to the baseline, but varying amounts of noise (mixed) lead to generally the best results. Too high amounts of noise (e.g. CER around 10 with WER approaching 70) hurts the supervised approach.